Tussen wiskunde en werkelijkheid

Tussen wiskunde en werkelijkheid

Inaugurele rede van Prof. Dr. Gerard J.P. van Breukelen,

benoemd tot hoogleraar Methodologie & Statistiek

in de Faculty of Health, Medicine & Life Sciences

en de Faculty of Psychology & Neuroscience,

uitgesproken op 17 april 2015 te Maastricht

Geachte rector magnificus en decanen, geachte collega's uit deze en andere universiteiten, lieve familieleden en vrienden. Het is een eer en genoegen u iets te mogen vertellen over mijn vakgebied, de Statistiek. Ik kwam erin terecht via omwegen. Eerst koos ik op school wat nu het Economie en Maatschappij profiel heet, om pas daarna te worden gegrepen door de schoonheid van de wiskunde. Vervolgens ging ik Psychologie studeren met uiteindelijk als hoofdvak Mathematische Psychologie en als bijvak wiskunde. Pas na mijn promotie ontdekte ik de clinical trials en steekproefberekeningen in de farmacie, en het observationele onderzoek en de multilevel designs in deze universiteit. Ik ben in de Statistiek dus een laatbloeier. De Statistiek is een deel van de Wiskunde en niet mediageniek. Maar het is een onmisbare pijler onder de mediagenieke wetenschappen economie, psychologie, geneeskunde. Er is geen Nobel prijs voor de wiskunde, maar zonder de wiskunde zouden er geen Nobel prijzen zijn.

De titel van mijn oratie is kort: "Tussen wiskunde en werkelijkheid". Als ondertitel voeg ik daar aan toe: "Het wordt tijd om Plato's grot te verlaten". Nu denkt u misschien dat ik bedoel dat statistici hun grot moeten verlaten om zich met de werkelijkheid bezig te houden. Dat bedoel ik echter niet, want de meeste statistici houden zich al met de werkelijkheid bezig. Statistiek is immers toepassingsgericht. Wat ik wel bedoel, is dat onderzoekers die hun data met statistische software analyseren, uit de grot moeten komen en de taal van de statistiek voldoende moeten leren om er verantwoord mee om te gaan. Sluipend dreigt er nl. een wiskundig analfabetisme. Jongeren zijn zeer vaardig met smartphones en internet, en er worden hoge eisen aan hun leertempo gesteld. De keerzijde is weinig tijd en geduld voor een technisch vak als de wiskunde, want dat leer je niet met zappen, scannen, en surfen. Tekenend zijn de matige rekenvaardigheid van jongeren, en het feit dat studenten bij het lezen van wetenschappelijke artikelen het onderdeel Methoden overslaan en voor de conclusies vertrouwen op de auteurs en reviewers. Die op hun beurt bij het lezen van andere artikelen overigens ook vaak de Methoden overslaan en vertrouwen op etc. In zo'n cultuur kan een simulatiestudie worden gepubliceerd om iets aan te tonen dat met 1 regel wiskunde kan worden bewezen (***zie slide 2 achterin dit bestand***). Die ene regel staat bovenin.

Een ander symptoom van wiskundig analfabetisme is het feit dat onze vakgroep in de laatste 10 jaren al haar promovendi op statistische projecten en al haar nieuwe docenten uit het buitenland heeft gehaald, vaak zelfs buiten Europa. Dit en de gestage groei van de mediagenieke wetenschappen waarin statistiek onmisbaar is, en de ontwikkeling van steeds complexere statistische methoden, nopen tot een goede statistische scholing van studenten en onderzoekers. Dat vraagt van hen abstractievermogen en wiskunde op het nivo van de middelbare school. En daar komt Plato's grotmetafoor om de hoek kijken. Deze wordt mooi geillustreerd door dit plaatje (***zie slide 3***). In Plato's metafoor zijn de mensen geketend aan de grond in een grot en zien zij het daglicht slechts indirect, via projecties van voorwerpen en gestalten op de muur. Zou een van hen vrij komen en de grot verlaten, dan zou die mens eerst verblind worden door het felle licht en slechts langzaam de echte voorwerpen gaan zien. Er zou tijd en oefening nodig zijn om te zien dat er in de grot slechts schaduwen zijn. Plato gebruikte deze metafoor om uit te leggen dat mensen slechts een zwak begrip hadden van abstracte vormen als "goedheid" en "vrijheid". De mens die de grot verliet stond voor de filosoof. Het was de taak van de filosoof om na het ontdekken van de werkelijkheid de nieuwe inzichten over te brengen op de medemensen in de grot, maar Plato waarschuwde dat dat moeilijk was.

Iets soortgelijks ervaart de statisticus in het onderwijs en de consultatie. De statisticus denkt in termen van abstracte vormen: designs, datastructuren, variabelen, functies. Studenten en onderzoekers vinden dat zelden leuk, want zij denken in termen van hun onderzoeksvraag. Hier zijn er een paar (***zie slide 4***). In elk voorbeeld wordt gevraagd naar het verschil in resultaat tussen twee situaties: een nieuwe behandeling versus geen of een oude behandeling. In statistische taal: we bekijken de relatie tussen twee variabelen: type behandeling en resultaat. Hiermee heb ik een eerste abstractie uit de statistiek genoemd: "**variabele**". Dat is een kenmerk waarop personen verschillen, dus iets dat tussen personen varieert. Niet elke leverpatient wordt op dezelfde wijze geopereerd, en niet elke leverpatient ligt even lang in het ziekenhuis. Dankzij die variatie tussen mensen kunnen we de relatie tussen variabelen onderzoeken: Zonder variatie geen correlatie.

Er zijn verschillende soorten variabelen. Lichaamsgewicht en reactietijd zijn **continu en kwantitatief**, want ze kennen een meeteenheid. Type behandeling en geslacht zijn **discreet of categorical**: ze leiden tot een categorisatie van mensen als: wel of niet behandeld, man of vrouw. Categorical variabelen met twee categorieen heten dichotoom. Met meer categorieen heten ze polytoom. Ook zijn er telvariabelen en overlevingsduren. Voor de eenvoud beperk ik me hier tot kwantitatieve en dichotome variabelen. Als we een relatie tussen twee variabelen causaal interpreteren, maken we nog een onderscheid. We noemen de oorzaak een **onafhankelijke variabele**, en het gevolg een **afhankelijke variabele**. In de getoonde voorbeelden is het type behandeling de onafhankelijke variabele en is het resultaat van de behandeling de afhankelijke variabele. Of we een relatie tussen behandeling en resultaat causaal mogen interpreteren hangt sterk af van het onderzoeksdesign en van de statistische analyse, zoals we nog zullen zien. Ook zullen we zien dat onderzoeken meestal meer dan twee variabelen kennen.

Een andere abstractie uit de statistiek is de **onderzoekseenheid** . In veel onderzoeken vormen personen de onderzoekseenheden, bijvoorbeeld: leerlingen, patienten, werknemers. We observeren of meten onze variabelen immers bij personen. Maar soms zijn de onderzoekseenheden organisaties, zoals scholen in een onderwijskundig onderzoek, of bedrijven in economisch onderzoek. En soms zijn de onderzoekseenheden delen van personen, zoals de hersenvoxels van een proefpersoon in een fMRI studie, of de genen van een patient. In veel onderzoeken zijn er **eenheden op meerdere nivos** , en is er bij elk nivo een onderzoeksvraag. Maar voorlopig gaan we uit van personen als eenheden.

Met deze twee abstracties, de "variabele" en "de "onderzoekseenheid", kunnen we ons verdiepen in de Statistiek. Graag zag ik dat u allen mijn oratie enigszins kunt volgen. Daarom begin ik bij de inleidende statistiek voor onze eerstejaars studenten, en bouw ik daarna op tot aan de statistiek voor promovendi. Daarbij beperk ik de wiskunde tot een minimum en toon ik vooral plaatjes. Ik excuseer me bij de collega die me vroeg om mijn rede te verluchtigen met grappen over Amsterdammers versus Limburgers. Dat gaat me niet lukken, want ik ben een Brabander. Voor een Limburger ben ik dus een Hollander, en voor een Hollander ben ik een Limburger. En daar hebben ze beiden wel een beetje gelijk in. Maar ik denk dat de Maagdenhuisbezetting veel zegt over het verschil tussen Amsterdam en Limburg. En laat ik het daar maar bij houden.

De Fundamenten

De eerste kennismaking van studenten met Statistiek is de frequentieverdeling van een variabele, aangevuld met samenvattende maten zoals gemiddelde en standaarddeviatie. De volgende figuur toont de scoreverdeling van een groep Nederlanders op de CFQ, een vragenlijst voor cognitive failures in het dagelijkse leven, zoals het vergeten van een afspraak (***zie slide 5***). Hoe hoger de score, hoe erger de failure. De meeste mensen scoren rond de 30, maar er is veel spreiding rond dat gemiddelde. De standaarddeviatie vat die spreiding samen. In het nieuws zien we regelmatig frequentieverdelingen, bijv. van inkomen of stemgedrag bij verkiezingen. Interessanter wordt het als we niet een enkele variabele beschrijven, maar de relatie tussen twee variabelen, zoals leeftijd en cognitieve uitval. In ons gegevensbestand vormen de personen de rijen, en vormen de variabelen de kolommen (***zie slide 6***). Elke rij bevat de waarden van een persoon op de beide variabelen (X = leeftijd, Y = CFQ score). Het verband tussen beide variabelen zien we met een spreidingsdiagram (***zie slide 7***). Daarbij zetten we de oorzaak op de horizontale X-as, en het gevolg op de verticale Y-as. Elke persoon is een punt. Overigens zijn deze data voor het doel van de oratie bewerkt, want in de echte data werd geen verband gevonden. De cognitieve uitval lijkt te stijgen met de leeftijd, maar het verband is zwak, want mensen met dezelfde leeftijd verschillen toch qua CFQ score. De relatie tussen leeftijd en CFQ score kunnen we samenvatten met een zo goed mogelijk passende lijn door de puntenwolk. Deze lijn kunnen we wiskundig beschrijven met de volgende functie (***zie formule in slide 7***):

$$Y = \beta_0 + \beta_1 X + e$$

Hierbij is Y de CFQ score en X de leeftijd. De grootheden β_0 en β_1 zijn **parameters** en karakteriseren het lineaire verband: β_0 is nu niet interessant , maar β_1 is de stijging van de CFQ score per jaar dat men ouder is. De term e van "error" geeft de verticale afstand aan tussen het punt en de lijn, en doet recht aan het feit dat de relatie tussen leeftijd en CFQ score zwak is: mensen met eenzelfde leeftijd verschillen onderling qua score. Dit heeft allerlei oorzaken, zoals genetische aanleg en opleidingsnivo. Het gaat ons nu om β_1 , de relatie tussen leeftijd en cognitieve uitval: Stijgt de uitval met het ouder worden, en zoja, hoe sterk is die stijging ? Met wiskunde kunnen we β_0 en β_1 berekenen, maar er is een probleem: We onderzoeken de relatie tussen leeftijd en score in de **populatie** van volwassen Nederlanders, maar door gebrek aan tijd en geld kunnen we ons onderzoek slechts uitvoeren in een klein deel van die populatie, een **steekproef** van bijv. 40 mensen zoals in de figuur. Zelfs als die steekproef willekeurig en daarmee representatief is, lopen we tegen het probleem van **steekproeftoeval** aan: Als we het onderzoek herhalen op een nieuwe steekproef, krijgen we een andere puntenwolk, andere lijn en andere waarde voor β_1 . Die waarde noteer ik als B1, want het is slechts een schatting die onderhevig is aan toeval. De volgende twee figuren tonen andere steekproeven uit dezelfde populatie (***slides 8 en 9***). Die toevalsvariatie tussen steekproeven wordt groter naarmate de punten verder van de lijn afliggen, immers: afwijkingen van de lijn geven CFQ verschillen aan tussen mensen die even oud zijn. Naarmate die verschillen groter zijn, maakt het voor de puntenwolk en de lijn meer uit welke mensen toevallig in de steekproef terecht komen. Gelukkig is er een remedie tegen grote steekproeffluctuaties,

nl. het **vergroten van de steekproef**: hoe meer personen we onderzoeken, hoe kleiner de fluctuatie wordt. De Maastricht Ageing study had dan ook een veel grotere steekproef dan 40 mensen. De statistiek voor eerstejaars studenten gaat vooral over de relatie tussen twee variabelen en de rol van steekproeftoeval. Met **statistische toetsen** kan men nagaan of een in de steekproef gevonden relatie betekent dat er in de populatie ook echt een relatie is, of dat de gevonden relatie mogelijk op toeval berust. Met **betrouwbaarheidsintervallen** kan men aangeven tussen welke grenzen de relatie in de populatie ligt, bijv. dat β_1 tussen 0.3 en 0.7 ligt als men in de steekproef 0.5 vindt. En bij de opzet van een nieuw onderzoek kunnen we berekenen hoe groot de steekproef moet zijn om een relatie aan te tonen als die echt bestaat. Voor een lineair verband toont de volgende figuur de vereiste steekproefomvang (***zie slide 10***). We zien dat die omvang groter wordt naarmate het aan te tonen verband zwakker is. Helaas kosten grote steekproeven veel geld en tijd, en zijn ze vaak niet haalbaar. Een belangrijk thema in onze vakgroep vormt dan ook de ontwikkeling van wiskundige formules en software om de beste steekproefomvang te berekenen voor complexe onderzoeksdesigns.

Vaak is de onafhankelijke variabele niet kwantitatief, maar dichotoom. Nemen we de relatie tussen CFQ score en geslacht in een kleine steekproef (***slide 11***). Op de X-as staan nu twee waarden: vrouw, man, met per waarde een kolom punten voor de CFQ scores van alle personen. De best passende lijn is de verbinding tussen de gemiddelde score van de mannen en die van de vrouwen. Steekproeftoeval leidt tot fluctuaties in elke kolom, en daarmee van beide gemiddelden en de helling van de lijn. Met de t-toets gaan we na of de gemiddelden genoeg van elkaar verschillen om te zeggen dat het geen toeval is. Meestal vervangt men deze figuur door een staafdiagram (***zie slide 12***). De hoogte van elke kolom geeft de gemiddelde CFQ score aan, en de antenne de spreiding. Deze figuur toont een belangrijk principe: De vraag naar het verband tussen geslacht en score kunnen we herformuleren als de vraag naar het verschil tussen mannen en vrouwen qua score. Omgekeerd kunnen we een verschil tussen groepen herformuleren als een relatie tussen variabelen. We kunnen dus kiezen tussen statistische methoden voor het verschil tussen groepen (t-toetsen en variantie analyse) en methoden voor het verband tussen variabelen (regressie analyse). Bij factoriele designs in de experimentele biologie en psychologie is ANOVA handiger. Bij observationele studies en clinical trials is regressie analyse echter beter.

Bekijken we tot slot de relatie tussen dichotome variabelen. Stel dat we cognitieve uitval dichotoom meten: wel uitval (1) of geen uitval (0). Dan zien we op elke as nog maar twee waarden (***zie slide 13***), en nog slechts vier punten, dus er liggen meerdere personen op eenzelfde punt. Om te zien hoeveel personen dat zijn, vervangen we elke punt door een getal dat het aantal personen aangeeft. Trekken we lijnen tussen die getallen, dan stappen we over van een spreidingsdiagram op een kruistabel (***slide 14***). Ook nu moeten we rekening houden met steekproeftoeval, en wel met de Chi-kwadraat toets.

Ons Platonische bouwwerk omvat al vele abstracties: Variabelen, onderzoekseenheden, correlatie, populatie, parameters, steekproef, schattingen, steekproeftoeval, steekproefgrootte, toets, betrouwbaarheidsinterval. Dit is het fundament. Noodzakelijk, maar nog niet zo spannend. Maar nu kunnen we de begane grond bouwen, en dan wordt het huis bewoonbaar. Die BG bestaat uit de toevoeging van variabelen, en drie abstracties: confounding, mediatie, en interactie (ook bekend als moderatie of effect modificatie).

De Begane Grond

Tot nu toe beperkte ik me tot twee variabelen, maar meestal zijn er meer, omdat de relatie tussen twee variabelen vaak afhangt van een derde (een moderator), of indirect verloopt via een derde (een mediator), of wordt vertekend door een derde (een confounder). We breiden ons databestand dus uit met een of meer kolommen (***zie slide 15***). Beginnen we met **confounding** of vertekening. Stel dat we nagaan of een therapie voor depressie effect heeft (***zie slide 16***). De onafhankelijke variabele X (therapie) heeft een nader te bepalen effect op de afhankelijke variabele Y (genezing). Maar een derde variabele (mate van depressie) heeft ook een effect op genezing, en hangt samen met therapie. We noteren de derde variabele met de C van confounder. Bekijken we een voorbeeld (***zie slide 17***). Op de X-as staat therapie. Op de Y-as staat genezing. We zien meer genezing in de therapiegroep dan in de controlegroep (paarse lijn). Maar laten we eens uitsplitsen naar de mate van depressie. We zien nu geen verschil in genezing meer tussen therapie en controle, noch bij de zware gevallen (rood), noch bij de lichte (blauw). Hoe kan dit ? De verklaring is simpel: de meeste lichte gevallen zoeken therapie, maar de meeste zware gevallen blijven thuis afwachten tot de bui overgaat. De therapiegroep bestaat dus vooral uit lichte gevallen en de controlegroep uit zware. Als we lichte en zware gevallen samenvoegen, eindigen we in de controlegroep nabij de rode lijn en in de therapiegroep nabij de blauwe lijn. En zoals de afstand tussen rode en blauwe lijn toont, is er bij de lichte gevallen meer genezing dan bij de zware, ongeacht therapie. De therapiegroep is dus al in het voordeel voordat er therapie wordt gegeven. Het omgekeerde kan ook (***slide 18***). Nu zien we geen therapie effect (paarse lijn). Maar als we uitsplitsen naar mate van depressie, zien we zowel bij de lichte gevallen (blauw) als de zware gevallen (rood) een stijgende lijn: meer genezing met therapie dan zonder therapie. De verklaring is zoals in de vorige figuur, maar andersom. Nu blijven de lichte gevallen thuis afwachten en gaan de zware gevallen in therapie, wellicht onder druk van hun partner. De therapiegroep start nu dus met een achterstand. Een therapie effect kan dus verdwijnen of verschijnen na correctie voor een confounder. Maar het kan nog gekker. Een averechts effect van therapie (paars) kan na correctie omkeren in een positief effect, of andersom (***slide 19***). Dit staat bekend als **Simpson's paradox** en kan optreden als de confounding heel sterk is.

De analyse van confounding is in de praktijk complexer dan dit. Om te beginnen moeten we de beide oorzaak-gevolg relaties, tussen therapie en genezing, en tussen mate van depressie en genezing, schatten met een steekproef, want we kunnen niet alle depressieve patienten in ons land onderzoeken. Onze resultaten lijden dus aan steekproeftoeval, en dat wordt erger naarmate de confounding sterker is. Maar bovendien kunnen er meerdere confounders zijn, zoals leeftijd, geslacht en opleidingsniveau. Dit is de bestaansreden voor regressie analyse. Hiermee kan men wiskundig corrigeren voor confounders zonder de steekproef op te splitsen. De volgende formule toont hoe dit werkt met 1 confounder, en de uitbreiding naar meerdere confounders is meer van hetzelfde (***zie formule en symbolen in slide 19***)**:**

$$Y = \beta_0 + \beta_1 X + \beta_2 C + e$$

Genezing (Y) is nu een functie van twee variabelen: therapie (X=0 voor controle, X=1 voor therapie)) en mate van depressie (C=0 voor zwaar en C=1 voor licht). Het effect van therapie na correctie voor de confounder is β_1. Dit is de helling van de rode en blauwe lijnen. Het effect van de confounder is β_2. Dit

is de afstand tussen de rode en blauwe lijn. In een regressie analyse passen we dit model toe op de onderzoeksgegevens. De analyse levert dan de beste schatting van de effecten van therapie en mate van depressie op de genezing, plus voor elk effect een significantietoets en een betrouwbaarheidsinterval.

Er is echter altijd een risico dat we confounders over het hoofd zien. Net als in de gezondheidszorg geldt in de statistiek daarom: Voorkomen is beter dan genezen. We kunnen confounding voorkomen door een sterk onderzoeksdesign te kiezen. We laten patienten niet zelf kiezen of ze in therapie gaan of niet, maar we laten het toeval bepalen in welke groep ze komen. We gooien bij elke patient een munt op: kop = therapie, munt = controle. Bij een voldoende grote steekproefomvang voorkomt dat confounding. In de Psychologie heet dit het **gerandomiseerde experiment**, maar ook wel **between-subject design**, omdat we het therapie effect schatten door een vergelijking te maken tussen twee groepen personen, de wel en niet behandelden. In de Geneeskunde heet dit de **randomised trial**, maar ook wel het **parallel groups design**. Uiteraard moeten we aan alle patienten eerst toestemming voor deelname vragen. Nog sterker is het **crossover design** zoals de Geneeskunde het noemt, of **within-subject design** zoals de Psychologie het noemt. Hierin is elke patient zijn of haar eigen controle, want we observeren elke patient zowel in de therapieconditie als in de controleconditie. Dit vereist een veel kleinere steekproef dan het parallel groups of between-subject design. Helaas kan het crossover design alleen gebruikt worden voor behandelingen waarvan het effect verdwijnt na stopzetten van de behandeling. Dus wel voor pijnstillers en fMRI experimenten, maar zelden voor chirurgie of psychotherapie.

Het gerandomiseerde experiment voorkomt dus confounding. Maar veel onderzoeksvragen lenen zich niet voor dit design, om ethische of logische redenen. Daarom zijn statistische regressiemethoden zo belangrijk. Hier zijn enige voorbeelden van niet-gerandomiseerde onderzoeken (***zie slide 20***). En zelfs in gerandomiseerde experimenten volstaat de eerstejaars statistiek zelden. Naast confounding bestaan er nl. nog twee fenomenen: mediatie en moderatie. Beginnen we met **mediatie (*zie slide 21*)**. Stel, we vergelijken psychotherapie met farmacotherapie voor depressie. De derde variabele is therapietrouw, de mate waarin men zich houdt aan de voorgeschreven behandeling. Dat is hier de mediator (M) of tussenschakel: De soort therapie die men krijgt heeft een effect op de therapietrouw, en die trouw heeft weer een effect op de genezing. Hierdoor krijgen we twee effecten van therapie op genezing: het directe effect (pad c) en het indirecte of gemedieerde effect (via pad a en pad b). Het totale therapie effect is de som van beide effecten. Om het **directe effect** van therapie op genezing te krijgen moeten we wel corrigeren voor de mediator. Maar om het **totale effect** van therapie op genezing te krijgen juist niet. Klassieke mediatie analyse bestaat daarom uit een combinatie van regressie analyses. Dit klinkt heel mooi, maar weer ligt confounding op de loer, en nu zelfs binnen het gerandomiseerde experiment ! We kunnen patienten wel willekeurig toewijzen aan een therapie, maar niet aan een mate van therapietrouw. Het effect van therapietrouw op genezing (pad b) kan dus worden vertekend door een vierde variabele C die met de mediator samenhangt en zelf een effect heeft op de genezing Y, bijvoorbeeld leeftijd. Als we niet corrigeren voor C, dan wordt het effect van de mediator op genezing (b) fout geschat. Daardoor wordt het indirecte effect van therapie op genezing (= a*b) fout geschat. En daardoor wordt ook het directe therapie effect (c) fout geschat. Slechts het totale therapie effect kan dankzij de randomisatie zonder vertekening worden geschat. Dit en diverse andere complicaties bij mediatie analyse hebben geleid tot een nieuw statistisch specialisme onder de naam “causal inference”.

We hebben nu confounding en mediatie gezien als rollen voor een derde variabele. Maar er is nog een rol, in de psychologie bekend als **interactie** of **moderatie**, en in de geneeskunde als **effect modificatie**. Dat ziet er zo uit (***zie slide 22***). Nu heeft de derde variabele, mate van depressie, *effect op de relatie* tussen therapie en genezing. De ernst van de depressie modereert, of modificeert, het effect van X op Y. Dit fenomeen kan zich voordoen in elk onderzoek, ook in een gerandomiseerd experiment, bijvoorbeeld aldus (***zie slide 23***). In de linkerplot is er een positief effect van therapie op genezing, maar sterker voor licht depressieve mensen (blauw) dan voor zwaar depressieven (rood). Houden we geen rekening met de mate van depressie, dan vinden we de paarse lijn. Die overschat het nut van therapie voor zware gevallen, en onderschat het voor lichte gevallen. In de rechterplot werkt de therapie goed voor de lichte gevallen (blauw), maar averechts voor de zware gevallen (rood). Negeren we de derde variabele, dan vinden we geen therapie effect (paars). Dat veel wetenschappers en politici geloven in moderatie blijkt uit het feit dat de EU miljoenensubsidies over heeft voor "personalised health care".

De klassieke methode om interactie te onderzoeken is variantie analyse (ANOVA). Dat is een "black box", maar als we die open maken zien we het volgende regressie model ***(zie slide 24):***

$$Y = \beta_0 + \beta_1 X + \beta_2 C + \beta_3 X C + e$$

Bovenop de effecten van therapie X en derde variabele C komt nu het effect van X*C. Dat we hiermee moderatie aankunnen, blijkt als we de vergelijking herschrijven:

$$Y = \left(\beta_0 + \beta_2 C \right) + \left(\beta_1 + \beta_3 C \right) X + e .$$

Het effect van therapie X op genezing hangt dus af van de derde variabele C.

Een bijzondere vorm van interactie is die van X met zichzelf. Dat levert dit model op:

$$Y = \beta_0 + \beta_1 X + \beta_2 X X + e ,$$ wat hetzelfde is als:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + e$$

Nu is Y geen lineaire functie van X meer, maar een kwadratische. Het effect van X op Y verandert nu mee met X zelf, en dit model kan vele effectpatronen aan, getuige de volgende figuur (***zie slide 25***). Voor dichotome onafhankelijke variabelen, zoals geslacht of therapie, is dit model niet nodig, want door twee punten past altijd een lijn. Het kwadratische model is echter nuttig voor kwantitatieve onafhankelijke variabelen, zoals leeftijd, behandelduur, of dosis medicatie. En als we het lineaire model loslaten, gaat de deur open naar vele modellen, zoals logarithmische en exponentiele. Bovendien kent bijna elk onderzoek meer dan drie variabelen. Statistisch modelleren gaat dan ook veel verder dan een gemiddelde, een correlatie, of een *p*-waarde. En hiermee is de BG van ons Platonische bouwwerk af. Gaan we nu naar de eerste verdieping, de multilevel designs.

De Verdieping

Ons bouwwerk begon met de abstracties "**variabele**" en "**onderzoekseenheid**". Deze zien we terug als kolommen en rijen in het databestand. We begonnen met drie kolommen (persoonsnr, X en Y). Daar voegden we kolommen aan toe voor confounders, mediators en moderators. We kunnen ook rijen toevoegen. Aan het begin van deze rede stelde ik dat de onderzoekseenheden meestal personen zijn, maar soms grotere eenheden zoals scholen of bedrijven. En soms kleinere, zoals herhaalde metingen, of genen, of voxels in hersenonderzoek. Richten we ons eerst op de grotere eenheden. Ons design wordt genesteld oftewel multilevel, want de personen zijn genesteld binnen organisaties, zodat we eenheden op twee nivos hebben, organisaties en personen. Ons databestand ziet er nu zo uit (***zie slide 26***). Voor de kolom met het persoonsnr staat een extra kolom met het organisatienr. We kunnen de datamatrix dus opdelen naar organisatie, en binnen elke organisatie hebben we een klassieke datamatrix.

We komen multilevel designs tegen in zowel experimenteel als observationeel onderzoek. Beginnen we met experimenten (***zie slide 27***). We kunnen organisaties toewijzen aan een van twee behandelingen, en geven alle personen in dezelfde organisatie dezelfde behandeling. Dat is een **"cluster randomised trial"**. We kunnen ook personen binnen organisaties toewijzen aan behandelingen, zodat binnen elke organisatie sommigen de ene behandeling krijgen en anderen de andere behandeling. Dat is een **"multicentre trial"** . Een multicentre trial staat tot een cluster randomised trial zoals een crossover trial tot een parallel groups trial, of zoals psychologen zeggen: zoals een within-subject design tot een between-subject design. Alleen zijn de personen nu vervangen door organisaties. In een multicentre trial is elke organisatie zijn eigen controle, en daardoor zijn voor een multicentre trial veel minder organisaties nodig dan voor een cluster randomised trial. Maar een multicentre trial is niet altijd haalbaar. Om te beginnen kan carry-over oftewel treatment contaminatie optreden, het doorlekken van een deel van de behandeling naar de controlegroep door communicatie binnen de organisatie. Dat risico is groter in de public health dan in de chirurgie. Maar bovendien is randomisatie van personen soms onmogelijk. Om na te gaan of een anti-pest programma op school werkt, kunnen we wel scholen of klassen toewijzen aan programma of controle, maar geen individuele leerlingen. Het programma wordt immers in de klas uitgevoerd. In onze universiteit worden veel cluster randomised en multicentre trials uitgevoerd. Hier zijn er een paar (***zie slide 28***). Doordat in multilevel designs niet alleen personen, maar ook organisaties onderling verschillen, ontstaat steekproeftoeval op twee nivos, cluster en persoon. Daar houdt multilevel analyse, in de biostatistiek bekend als mixed regression, rekening mee. Een analyse die de clustering negeert, onderschat het steekproeftoeval en leidt te snel tot de conclusie dat de behandeling werkt.

Bij cluster randomised en multicentre trials speelt confounding geen rol dankzij de randomisatie. Maar multilevel designs komen ook in niet-gerandomiseerd onderzoek voor, en dan is confounding weer een probleem. Neem bijv. de relatie tussen studie inzet en examenuitslag bij 15 leerlingen in 3 scholen (***zie slide 29***). We zien in elke school een positief verband tussen inzet en examenuitslag: hoe meer men studeert, hoe beter het resultaat (grijze lijnen). Maar als we de gemiddelde inzet en uitslag per school bekijken, zien we een negatief verband: hoe meer men studeert, hoe slechter men presteert (rode lijn). Niet meer studeren dus? Deze tegenstrijdigheid staat bekend als "ecological fallacy", of als "aggregation bias", of als **"Robinson's paradox"**. Het is een vorm van confounding, met de school als confounder. In

de upper class school zijn de leerlingen luier dan in de lower class school, maar bij gelijke inzet presteert een leerling in de upper class school beter dan in de lower class school. Dat kan vele oorzaken hebben. Misschien heeft de upper class school betere leraren. Als het onderzoek 3 scholen telt, kunnen we school als confounder meenemen. Voor 30 scholen is echter multilevel analyse nodig met een wiskundig model dat ik u bespaar. Wie deze methode niet kent, zal ofwel de data aggregeren tot het schoolnivo (rode lijn), ofwel disaggregeren tot het leerlingnivo zonder onderscheid naar school. En dan vinden we helemaal geen verband meer (blauwe lijn), doordat we twee verbanden op een hoop gooien: het positieve binnen scholen, en het negatieve tussen scholen. Robinson's paradox ligt ook op de loer in gezondheidsonderzoek. Een ziekenhuis kan de zorgkwaliteit verhogen met slim beleid, zoals een checklist bij de overdracht van patienten tussen afdelingen. Toch kunnen ziekenhuizen met een slim beleid slechter scoren op ligduur of mortaliteit dan ziekenhuizen zonder dat beleid. Ziekenhuizen verschillen immers op vele variabelen, zoals de gezondheid van de bevolking in die regio. De relatie tussen beleid en resultaat kan over ziekenhuizen heen bezien dus heel anders zijn dan de relatie binnen ziekenhuizen. De volgende figuur toont Robinson's paradox zelfs in een psychologisch experiment (***zie slide 30***). In plaats van scholen en leerlingen hebben we personen en test items. We onderzoeken de relatie tussen reactietijd en juistheid van antwoorden op bijv. rekensommen. Bij elke persoon zien we wat we verwachten: als men meer tijd besteedt, is het antwoord beter. Maar wat als we per persoon de gemiddelde reactietijd berekenen en het percentage goede antwoorden, en die twee correleren ? Dan vinden we een negatief verband. Moeten we dus snel rekenen om goed te scoren ? Natuurlijk niet. Personen verschillen in aanleg voor, en ervaring met, rekenen. Daardoor kan de een sneller en toch nauwkeuriger rekenen dan de ander. Correlaties over personen heen berekend, zeggen dus niets over correlaties binnen personen.

Zoals dit voorbeeld toont, is er niet alleen een multilevel design als we personen binnen organisaties onderzoeken, maar ook bij herhaalde metingen binnen personen. Neem de tinnitus trial. Hierin werd cognitieve gedragstherapie vergeleken met de gebruikelijke zorg. Een van de uitkomstmaten was de ernst van de tinnitus, gemeten bij aanvang van de behandeling en na 3, 8 en 12 maanden. De volgende figuur toont het beloop over de tijd (***slide 31***). We zien dat het verschil tussen de twee groepen toeneemt ten gunste van de cognitieve therapie, en dit verschil is significant. De data hadden geanalyseerd kunnen worden met herhaalde metingen ANOVA. Maar ze zijn geanalyseerd met multilevel regressie, met patient en meting als nivo's, omdat men dan patienten met een ontbrekende meting kan meenemen. Multilevel analyse is echter ook nuttig in laboratorium experimenten waarin personen reageren op een reeks stimuli, zoals plaatjes of rekensommen, en we de effecten van stimuluskenmerken op die reacties onderzoeken. Met multilevel analyse zijn single trial analyses mogelijk. Hier zijn een paar voorbeelden (***slide 32***). Het feit dat clinical trials en psychologische experimenten met zowel herhaalde metingen ANOVA als multilevel regressie kunnen worden geanalyseerd, wijst op een speciale status van herhaalde metingen. In ANOVA worden de vier metingen in de tinnitus trial opgevat als vier variabelen, dus als kolommen in onze datamatrix. Dit zgn. multivariate dataformat ziet er aldus uit (***slide 33***). In multilevel analyse echter worden de vier metingen opgevat als onderzoekseenheden die genesteld zijn binnen proefpersonen, dus als rijen in onze datamatrix. Dit zgn. univariate dataformat ziet er zo uit (***slide 34***). We kunnen herhaalde metingen dus zien als extra variabelen of als extra onderzoekseenheden. Soms is het eerste handiger, vooral in

laboratorium experimenten met een factorieel design. Maar soms is multilevel analyse nodig, vooral buiten het lab bij missing data, en in het lab bij single trial analyses. Het onderscheid tussen variabelen (kolommen) en eenheden (rijen) is hier dus niet houdbaar. Dit duale karakter van herhaalde metingen in de statistiek doet denken aan dat van licht in de natuurkunde: soms is het handiger om licht op te vatten als een golf, soms als een deeltje.

Dit is een goed moment voor een laatste paradox, want de oplossing ervan dank ik aan het duale karakter van herhaalde metingen. De volgende figuur toont **Lord's paradox**, die op de loer ligt als men groepen vergelijkt die niet via randomisatie zijn gevormd en al verschillen voordat de behandeling start (***slide 35***). De figuur betreft het effect van een preventieprogramma voor adolescenten met een verhoogd risico op depressie. Al voor aanvang van het programma is er een significant verschil: de controlegroep vertoont minder symptomen dan de programmagroep. We kunnen hier met twee methoden voor corrigeren. Methode 1 berekent eerst per persoon de verandering in symptomen tussen voor- en nameting, en vergelijkt dan de groepen op gemiddelde verandering (CHANGE). Dat levert een significant verschil op: in de programmagroep dalen de symptomen meer dan in de controlegroep. Het programma werkt dus. Methode 2 vergelijkt de groepen op de nameting, en neemt de voormeting mee als confounder (ANCOVA). Dat levert geen verschil op. Het programma werkt dus niet. Maar wat is de juiste conclusie ? Hier wijst methode 1 dus wel op een effect, maar methode 2 niet. Het omgekeerde komt ook voor. Stel dat de depressiestudie het volgende plaatje had opgeleverd (***slide 36***). Nu is er geen effect volgens methode 1 (CHANGE), maar wel volgens methode 2 (ANCOVA). Wat is nu de juiste conclusie ? Dit probleem hield mij jaren bezig en het lezen van tientallen artikelen gaf me geen antwoord. Ik ben oud-collega Hubert Schouten erkentelijk voor het attenderen op een artikel van Laird en Wang dat de sleutel tot de oplossing gaf. Daarin werd vermeld dat ANCOVA equivalent is aan een multilevel model voor herhaalde metingen waarin de groepen niet verschillen op de voormeting. Dat zou de verklaring zijn, immers: bij Lord's paradox is er altijd een groepsverschil op de voormeting. Toepassing van dat multilevel model op diverse data bevestigde wat Laird en Wang stelden, maar helaas gaven zij geen wiskundig bewijs en ook geen verwijzing. Omdat ik dat bewijs ook nergens kon vinden, heb ik het zelf maar geleverd. Het resultaat is in 2013 gepubliceerd. De oplossing van Lord's paradox zit dus in het duale karakter van herhaalde metingen. Nergens in het ANCOVA model, waar de voormeting een andere variabele is dan de nameting, is te zien dat ANCOVA aanneemt dat de groepen niet verschillen op de voormeting. Pas na herschrijving naar een multilevel model waarin de voormeting een andere eenheid is dan de nameting, wordt de ANCOVA aanname zichtbaar die Lord's paradox verklaart.

Dit betekent overigens niet dat de CHANGE methode de juiste methode is, want ook die maakt een aanname, nl. dat twee groepen zonder behandeling even sterk zouden veranderen tussen voor- en nameting. Het feit dat de groepen op de voormeting verschillen, doet vermoeden dat ze zich in het verleden verschillend hebben ontwikkeld, en dat maakt de aanname onder CHANGE twijfelachtig. Gelukkig is er een praktisch advies voor onderzoekers die niet-gerandomiseerde groepen vergelijken: Pas beide methoden toe, zowel ANCOVA als CHANGE. Als ze elkaar niet tegenspreken qua conclusie, maar slechts verschillen in effect grootte, zit men vermoedelijk goed. Beide methoden zijn nl. de uitersten van een algemenere methode die niet toepasbaar is om redenen die ik u bespaar. In de hier getoonde depressiestudie spraken de methoden elkaar wel tegen, en was geen conclusie mogelijk.

Hiermee is de verdieping van ons Platonische gebouw voltooid. Ik zou nog een zolder kunnen toevoegen met speciale onderwerpen, zoals de analyse van meetinstrumenten, hoogdimensionele data, Bayesiaanse statistiek, en missing data. Daar is echter geen tijd voor, en bij elk onderwerp ken ik collega's die er beter over kunnen vertellen dan ik. Het is daarom hoog tijd om iets te zeggen over de plaats van onze vakgroep in Plato's gebouw, en daarna af te sluiten.

Ons eigen onderzoek: toen, nu, straks

Toen ik 24 jaar geleden naar deze universiteit kwam, had de vakgroep geen onderzoeksprogramma en deed slechts een enkele statisticus eigen onderzoek naast het onderwijs en de consultatie. Ik deed zelf nog wat onderzoek naar psychometrische reactietijdmodellen als vervolg op mijn promotie, en ontdekte dat psychometrische modellen multilevel modellen waren. Intussen was ik via de consultatie bekend geraakt met multilevel analyse van cluster randomised trials. Op dat moment trad mijn voorganger Martijn Berger als hoogleraar aan. Hij wilde een onderzoeksprogramma starten rond Optimal Design en gaf een voorbeeld uit de psychometrie. Dat sprak me niet zo aan, want ik wilde de psychometrie verlaten voor de statistiek van multilevel trials. Gelukkig bleek ook dat thema zich te lenen voor Optimal Design, en zo was het eerste promotieproject geboren: Optimale steekproefgroottes voor cluster randomised en multicentre trials. Daarna volgden projecten over optimal design van kennistoetsen, clinical trials met herhaalde metingen, cohort studies, en fMRI experimenten (***slide 37***). Bij sommige projecten was ik copromotor, en daar heb ik zelf veel van geleerd. Nu hebben we nog twee projecten lopen over multilevel experimenten, en binnenkort starten we er een over multilevel surveys, geinspireerd door expertise werk voor de European Food Safety Authority en het Nederlands Instituut voor Psychologen. Het Optimal Design onderzoek heeft onze vakgroep veel nuttige kennis opgeleverd voor de berekening van steekproefgrootten voor cluster randomised trials en de bepaling van het beste aantal meetmomenten in een cohort studie. Ik ben Martijn erkentelijk voor het initieren van dit onderzoek, waarvan ik nu een paar resultaten toon.

In een cluster randomised trial worden complete organisaties, zoals scholen of huisartspraktijken, aan behandeling of controle toegewezen. Een belangrijke vraag bij de planning is de steekproefomvang: Hoeveel clusters zijn nodig, en hoeveel personen per cluster ? De volgende figuur toont het optimale aantal clusters en personen per cluster, als functie van de intraclasse correlatie of ICC (***slide 38***). Die ICC geeft aan welk deel van de uitkomst variatie tussen clusters is i.p.v. tussen personen. Hoe groter de ICC, hoe meer steekproeftoeval op het clusternivo, en dus hoe meer clusters er nodig zijn, en hoe minder personen per cluster (dat laatste onder aanname van een vast budget). Als dit optimale design onvoldoende "power" levert om een behandeleffect aan te tonen, moet het budget omhoog en dat extra geld moet gestoken worden in meer clusters, niet in meer personen. Analoge resultaten gelden voor multicentre trials. Een ander resultaat is het optimale aantal metingen in een clinical trial met een gegeven follow-up duur per patient. De belangrijkste bevindingen staan in deze tabel (***slide 39***). Het optimale aantal metingen hangt af van het correlatiepatroon van de metingen, van het verloop van het behandel effect over de tijd, en van de kosten per meting en per patient. Maar vaak volstaan 3 metingen per patient: 1 aan het begin, 1 halverwege, en 1 aan het einde. Als er veel dropout wordt verwacht, kan men beter 5 metingen plannen. Maar in sommige trials worden 6 tot 12 metingen verricht. Er valt dus nog wat te besparen.

De laatste jaren is het onderzoek in onze vakgroep verbreed naar diverse methoden voor data analyse. Er loopt onderzoek naar multivariate interrater agreement in een multilevel situatie (van belang voor de diagnostiek in de zorg en examinering in het onderwijs). Er loopt onderzoek naar growth curve modeling (van belang voor onderzoek naar veroudering en lifetstyle verandering). En er loopt onderzoek naar het ontdekken van interacties tussen personen en situaties (van belang voor personalised health care). Met de komst van twee nieuwe statistici medio dit jaar gaat er onderzoek lopen naar multiple imputatie methoden voor missing data (vooral van belang buiten het laboratorium), en Bayesiaanse methoden voor hoogdimensionele data (vooral van belang in het laboratorium). Zelf wil ik na het afronden van enige artikelen me deels richten op "causal inference". Toepassingen zie ik overal waar men causale ketens veronderstelt. Maar ook het optimal design onderzoek gaat door, zij het op een iets lager pitje.

Het is tijd om af te ronden. Ik hoop u een kijkje te hebben gegeven in de wereld van de statistiek, zeg maar de wereld buiten Plato's grot. De schaduwen in het verhaal van de onderzoeker die om statistische hulp vraagt, vertaal ik in Platonische vormen: Wat zijn de variabelen ? Hoe zijn hun onderlinge relaties ? Is er sprake van een multilevel design ? In die wereld van wiskundige abstracties zie ik beter wat het design is en hoe het geanalyseerd moet worden, dan in de grot. Als ik dan een antwoord heb gevonden op de vraag van de onderzoeker, keer ik terug in de grot en vertaal mijn antwoord in termen van de schaduwen die de onderzoeker ziet. Die schaduwen zijn niet minder belangrijk door het feit dat het schaduwen zijn. Onze wereld bestaat immers uit schaduwen. En uiteindelijk dient de wiskunde niet zichzelf, maar andere wetenschappen en daarmee de samenleving. Maar ik vind het heerlijk om af en toe uit de grot te treden en me met wiskunde bezig te houden, net zoals het heerlijk is om af en toe boven de wolken te vliegen en de aarde vanuit een ander perspectief te zien. Ook denk ik dat het de taak van statistici is om studenten in het onderwijs, en onderzoekers in de consultatie, de grot uit te leiden en zelfredzamer te maken als consumenten en producenten van wetenschappelijk onderzoek. Het volstaat voor een consument niet om in artikelen de samenvatting en discussie te lezen, en voor de rest te vertrouwen op de schrijver. Het volstaat voor een producent niet om software aan te sturen en de p-waarde op te zoeken. Slechts wie in staat is eenvoudige wiskundige modellen te begrijpen, kan onderzoeksdesigns en statistische analyses begrijpen. Wie geen wiskunde beheerst, is een analfabeet in die wetenschappen die zwaar leunen op de statistiek. En er zijn te weinig statistici om alle analfabeten voor te lezen. We moeten de analfabeten dus leren zelf te lezen. Dat vraagt inspanningen van alle partijen. Maar uiteindelijk verhoogt dat de kwaliteit van de wetenschap en van onze samenleving.

Rest mij het dankwoord. Allereerst aan de besturen van deze universiteit, van de Faculteit Health, Medicine and Life Sciences, en van de Faculteit Psychology & Neuroscience, voor het vertrouwen dat ze in me hebben gesteld door mij te benoemen als hoogleraar, en door in te stemmen met de instelling van de interfacultaire vakgroep Methodologie & Statistiek. Ik doe mijn best om uw vertrouwen waardig te zijn door de vakgroep zo te leiden dat zowel de beide faculteiten als de vakgroepsleden daar tevreden mee zijn. Dat gaat met vallen en opstaan, maar zo leert een kind lopen. Dank ook aan mijn voorganger Martijn Berger voor het leiden van onze vakgroep gedurende 16 jaren. Ook na je emeritaat ging je door met het begeleiden van promovendi, en dit jaar gaat er nog een oud-collega bij je promoveren. Dank ook aan Frederik-Jan van Schooten voor het leiden van onze vakgroep in het eerste jaar na Martijn's emeritaat. Dat was voor de vakgroep geen gemakkelijk jaar en jij kreeg veel op je af, terwijl je werk bij

Toxicologie doorging. Ook dank ik al mijn collega's en oud-collega's voor de goede samenwerking, en voor het geduld met mijn ongeduld. Met name dank ik Math, Frans, Ton, en Marga. Marga, als ik zeg dat een hoogleraar best een maandje gemist kan worden, maar een secretaresse nog geen week, dan meen ik dat. Ook mijn dank aan alle promovendi, oud-promovendi, studenten en oud-studenten, voor de vele vragen die me ertoe dwongen om dieper na te denken of beter te formuleren. Dank ook aan mijn ouders en schoonouders voor vele jaren van liefde, zorg en gastvrijheid. Hoe jammer dat jullie er niet bij kunnen zijn. Dank aan alle aanwezigen. Fijn dat jullie er wel bij zijn. En tot slot mijn gezin. Aline en Julie, de dagen waarop jullie werden geboren waren de mooiste dagen van mijn leven. Hoe zeer ik ook houd van de Statistiek, nog meer houd ik van jullie. Wees niet bang om fouten te maken, want daar leer je van. Ik kan het weten, want ik maak nog steeds fouten. Degene die dat het beste weet, is mijn echtgenote. Daniella, wij delen het leven nu al ruim 30 jaar en dat was geen fout. Ons leven gaat niet over rozen, want daarvoor is ons arbeidsethos te groot. Gelukkig delen we met elkaar en de kinderen onze liefde voor eenvoud, een goed boek, en vakantie op een eiland. Er zijn spannendere huwelijken dan het onze. Maar dat we al 30 jaar samen zijn, zegt genoeg. Ik hoop dat we daar nog 30 goede jaren aan toe kunnen voegen. Je had enige bedenkingen bij mijn hoogleraarschap, want we hadden al zoveel werk. Toch heb je mijn keuze gerespecteerd en gefaciliteerd, en dat ik hier nu sta is ook jouw verdienste. Mijn dank, respect en liefde zijn voor jou. Straks drinken we samen een glas, en morgen kijken we misschien weer Downton Abbey. Soms is het leven simpel.

Ik heb gezegd (***slide 40***)

Maastricht University Leading in Learning!

Tussen Wiskunde en Werkelijkheid

Maastricht University Leading in Learning!

$$Y = a + bX \quad \Rightarrow \quad Y - X = a + (b-1)X$$

International Journal of Psychophysiology 52 (2004) 277–283

Analysis of covariance (ANCOVA) with difference scores

John Jamieson*

Department of Psychology, Lakehead University, Thunder Bay, ON, Canada P7B 5E1

Received 3 June 2003; received in revised form 4 December 2003; accepted 4 December 2003

Abstract

When comparing pretest to posttest changes in non-randomized groups, most researchers are correctly avoiding ANCOVA with posttest as the dependent variable and pretest as the covariate. However, there is a widespread use of ANCOVA in which the difference score (posttest minus pretest) is used as the dependent variable, and pretest as the covariate. A computer simulation study is presented which shows that measurement error causes identical, biased conclusions when comparing changes using either the posttest score or the posttest minus pretest difference score as the dependent variable. The reasons for this bias are explained and illustrated.

Keywords: Analysis of covariance (ANCOVA); Difference scores; Change; Measurement error

Maastricht University Leading in Learning!

Bron: http://www.dailyimpact.net//?s=plato%27s+cave

Maastricht University Leading in Learning!

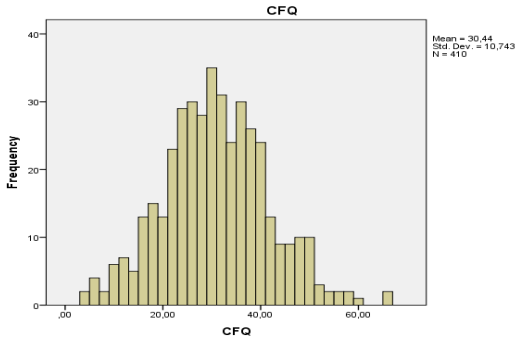
Wat is het effect van een stress management lesprogramma op de stress, coping, angst en depressie bij kinderen in het basisonderwijs ?
(Kraag et al., Journal of Child Psychology and Psychiatry, 2009)

Effecten van gespecialiseerde zorg op basis van cognitieve gedragstherapie versus gebruikelijke zorg op de kwaliteit van leven en ernst van klachten bij tinnituspatienten (Cima et al., The Lancet, 2012)

Effect van open versus laparoscopische leverchirurgie op functioneel herstel en duur van de ziekenhuisopname (Van Dam et al., BMC Trials, 2012).

Effecten van bewonersgerichte versus traditionele zorg in verpleeghuizen op de werkomstandigheden van de verzorgers (Berkhout et al., Work & Stress, 2003)

Maastricht University Leading in Learning!



Bron: Maastricht Ageing study (deel van het totale panel, 1e follow-up)

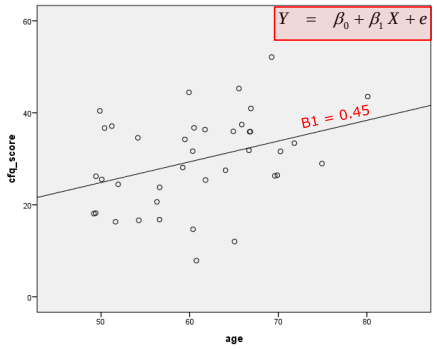
FHML & FPN

Maastricht University Leading in Learning!

persoonsnummer	leeftijd	CFQ score
1	51	8
2	57	20
3	68	34
4	62	19
5	74	53
6	56	27
7	80	44
8	65	49
Etc.		

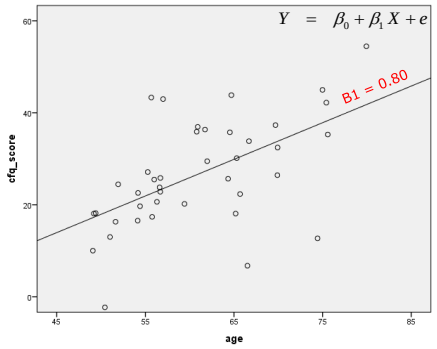
FHML & FPN

Maastricht University Leading in Learning!



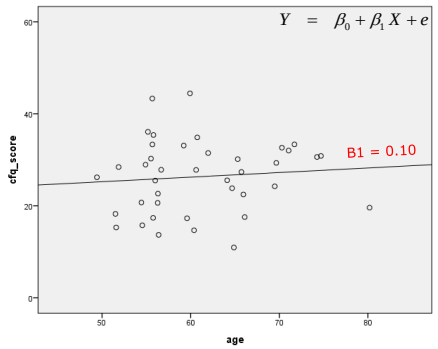
FHML & FPN

Maastricht University Leading in Learning!



FHML & FPN

Maastricht University Leading in Learning!

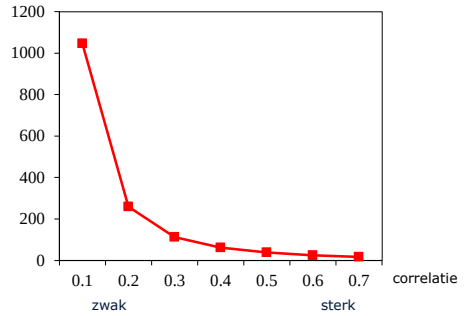


FHML & FPN

9

Maastricht University Leading in Learning!

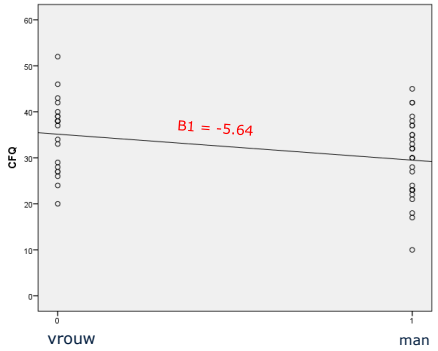
vereist aantal personen
voor aantonen correlatie
(α= 5%, power 90%)



FHML & FPN

10

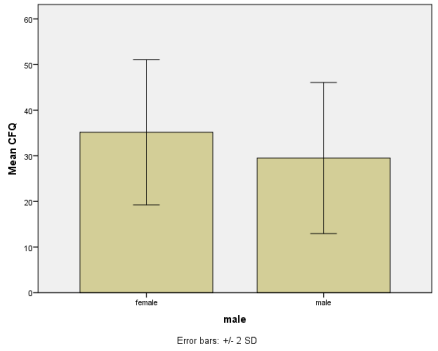
Maastricht University Leading in Learning!



FHML & FPN

11

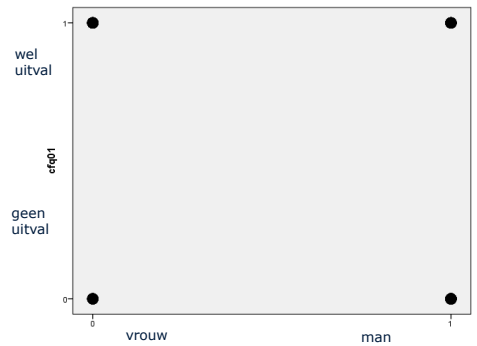
Maastricht University Leading in Learning!



FHML & FPN

12

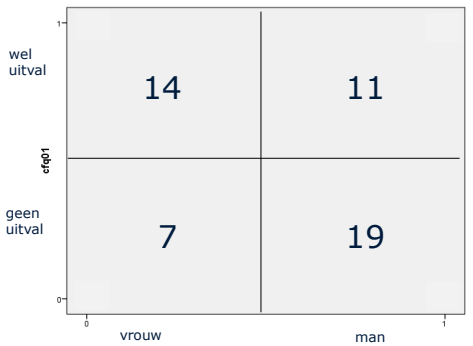
Maastricht University Leading in Learning!



FHML & FPN

13

Maastricht University Leading in Learning!



FHML & FPN

14

Maastricht University Leading in Learning!

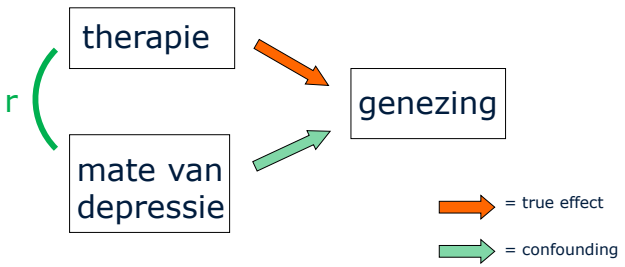
Persoons nummer	X (oorzaak)	Y (gevolg)	Confounder	Mediator	Moderator
1					
2					
3					
4					
5					
6					
7					
8					
Etc.					

FHML & FPN

15

Maastricht University Leading in Learning!

Confounding

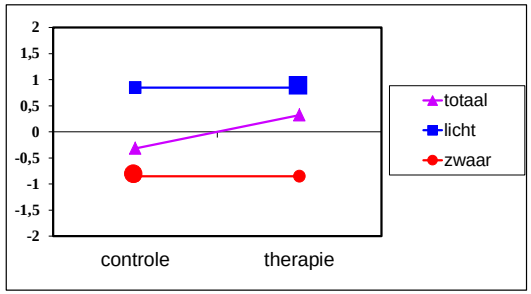


FHML & FPN

16

Maastricht University Leading in Learning!

logodds genezing

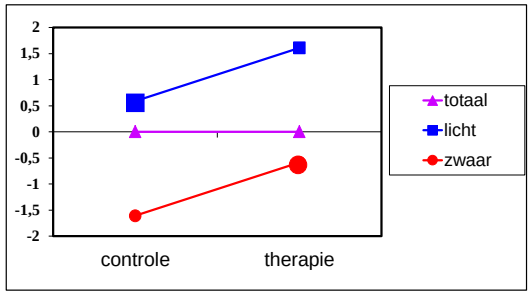


FHML & FPN

17

Maastricht University Leading in Learning!

logodds genezing



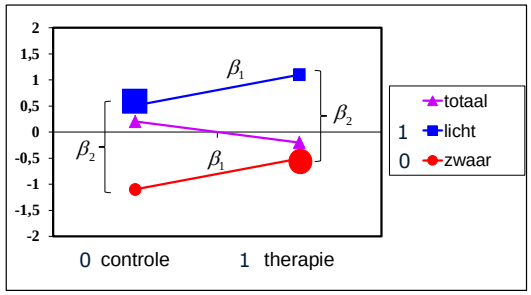
FHML & FPN

18

Maastricht University Leading in Learning!

$$Y = \beta_0 + \beta_1 X + \beta_2 C + e$$

logodds genezing



FHML & FPN

19

Maastricht University Leading in Learning!

Huver et al. (2007). Parenting styles and adolescent smoking cognitions and behavior. *Psychology and Health*, *22*, 575-593.

Van Breukelen & Vlaeyen (2005). Norming clinical questionnaires with multiple regression: the pain cognition list. *Psychological Assessment, 17,* 336-344.

Appels et al. (2000). Behavioral risk factors of sudden cardiac arrest. *Journal of Psychosomatic Research*, *48*, 463-469.

Engels et al. (1999). Influences of parental and best friends' smoking and drinking on adolescent use: a longitudinal study. *Journal of Applied Social Psychology*, *29*, 337-361.

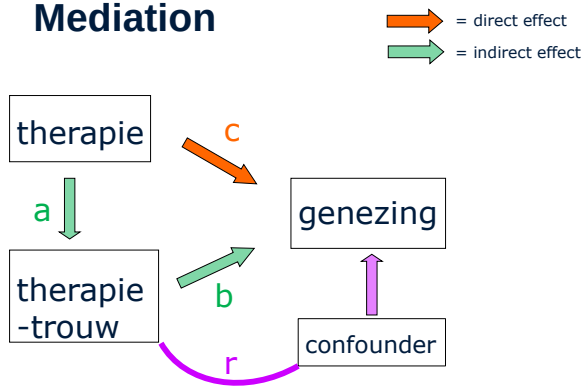
De Jonge et al. (1996). Testing the demand-control-support model among health-care professionals: a structural equation model. *Work and Stress*, *10*, 209-224.

FHML & FPN

20

Maastricht University

Mediation

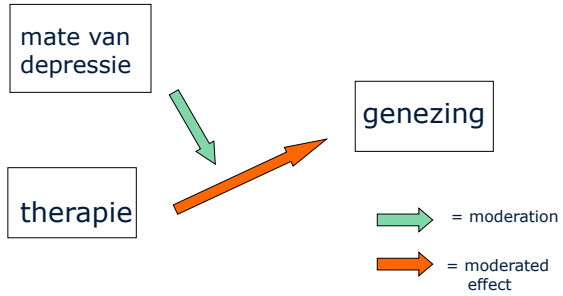


FHML & FPN

21

Maastricht University

Moderation (interaction)

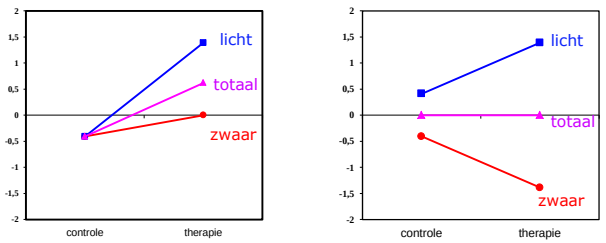


FHML & FPN

22

Maastricht University

logodds
genezing



FHML & FPN

23

Maastricht University

Regressiemodel met interactie

$$Y = \beta_0 + \beta_1 X + \beta_2 C + \beta_3 X C + e$$

$$Y = (\beta_0 + \beta_2 C) + (\beta_1 + \beta_3 C) X + e$$

Interactie van X met zichzelf

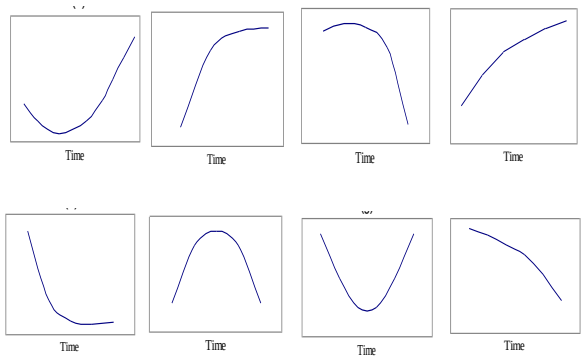
$$Y = \beta_0 + \beta_1 X + \beta_2 X X + e$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + e$$

FHML & FPN

24

Maastricht University Leading in Learning!



Bron: PhD thesis Abebe, 2014

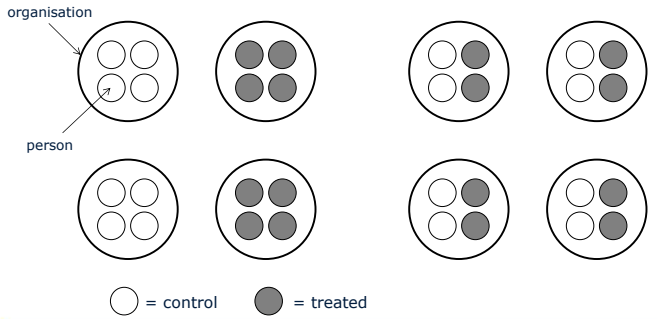
Maastricht University Leading in Learning!

organisatie	persoon	X (oorzaak)	Y (gevolg)	confounder	Mediator	Moderator
1	1					
1	2					
1	3					
1	4					
2	5					
2	6					
2	7					
2	8					
Etc.	Etc.					

Maastricht University Leading in Learning!

cluster randomized trial

multicentre trial



Maastricht University Leading in Learning!

***Cluster randomised trials*:**

Slok et al. (2014). Effectiveness of the Assessment of Burden of Chronic obstructive pulmonary disease (ABC) tool: study protocol of a cluster randomised trial in primary and secondary care. *BMC Pulmonary Medicine.*

Kraag et al. (2009). 'Learn young, learn fair', a stress-management programme for 5th and 6th graders: longitudinal results from an experimental study. *Journal of Child Psychology and Psychiatry*.

Steenhuis et al. (2004). The effectiveness of nutrition education and labeling in Dutch supermarkets. *American Journal of Health Promotion.*

Ausems et al. (2002). Short-term effects of a randomised computer-based out-of-school smoking prevention trial aimed at Dutch elementary schoolchildren. *Preventive Medicine.*

Multicentre trials:

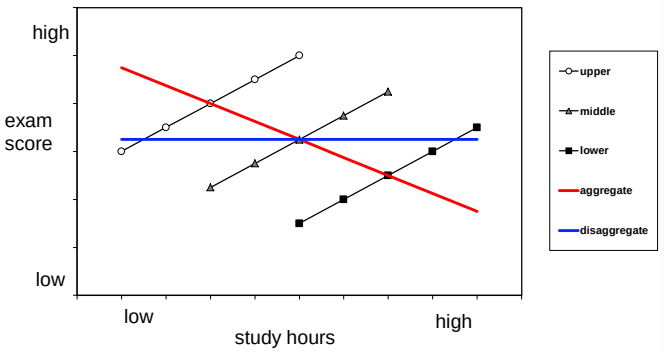
Wetzelaer et al. (2014). Design of an international multicentre RCT on group schema therapy for borderline personality disorder. *BMC Psychiatry*.

Van Dam et al. (2012). Open versus laparoscopic left lateral hepatic sectionectomy within an enhanced recovery ERAS(R) program (ORANGE II-Trial):study protocol for a randomized controlled trial. *Trials.*

Van Keulen et al. (2011). Tailored print communication and telephone motivational interviewing are equally succesful in improving multiple lifestyle behaviors in an RCT. *Annals of Behavioral Medicine.*

Leeuw et al. (2008). Exposure in vivo versus operant graded activity in chronic low back pain patients: results of a randomized controlled trial. *Pain.*

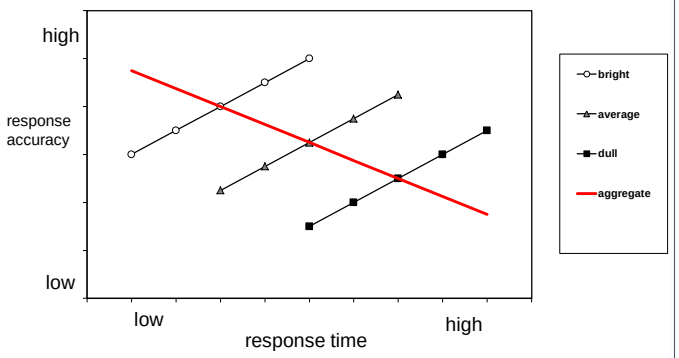
Maastricht University Leading in Learning!



FHML & FPN

29

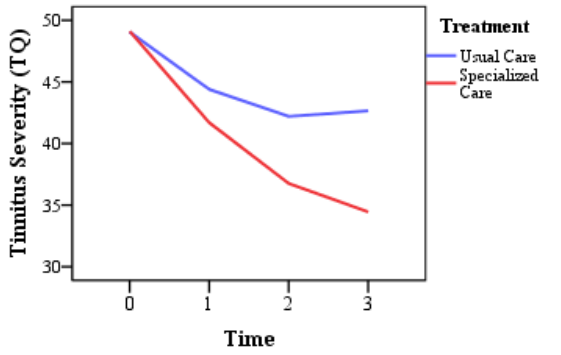
Maastricht University Leading in Learning!



FHML & FPN

30

Maastricht University Leading in Learning!



bron: PhD thesis Cima, 2013

FHML & FPN

31

Maastricht University Leading in Learning!

Single trial analyses met multilevel (mixed) regressie :

Vossen et al. (2011). Association between event-related potentials and pain ratings: not as straightforward as often thought. *Journal of Psychophysiology*.

Nentjes et al. (2015). Examining the influence of psychopathy, hostility biases, and automatic processing on criminal offenders' Theory of Mind. *International Journal of Law and Psychiatry*.

De Kinderen et al. (2015). From clinically relevant outcome measures to quality of life in epilepsy: a time trade-off study (submitted).

Van Breukelen (2005). Psychometric modeling of response speed and accuracy with mixed and conditional regression. *Psychometrika*.

FHML & FPN

32

Maastricht University Leading in Learning!

Multivariaat dataformat

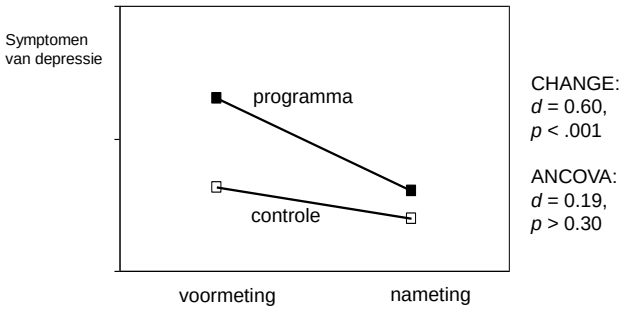
patient	therapie	Tinnitus1	Tinnitus2	Tinnitus3	Tinnitus4
1	0				
2	1				
3	0				
4	1				
5	0				
6	1				
Etc.					

Maastricht University Leading in Learning!

Univariaat dataformat

patient	therapie	meetmoment	Tinnitus
1	0	1	
1	0	2	
1	0	3	
1	0	4	
2	1	1	
2	1	2	
2	1	3	
2	1	4	
Etc.			

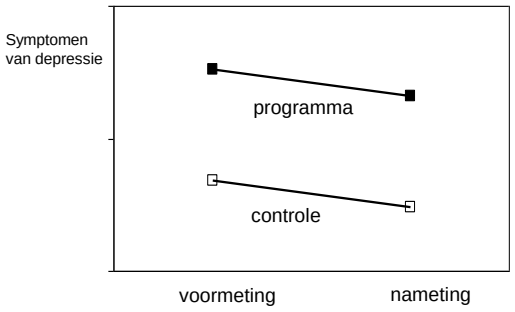
Maastricht University Leading in Learning!



CHANGE:
$d = 0.60$,
$p < .001$

ANCOVA:
$d = 0.19$,
$p > 0.30$

Maastricht University Leading in Learning!



CHANGE:
$d = 0.01$,
$p > 0.90$

ANCOVA:
$d = -0.28$,
$p < 0.01$

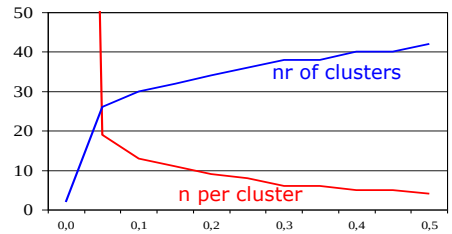
Maastricht University Leading in Learning!

PH theses Optimal design dept. M&S

Maastricht University Leading in Learning!

Optimal cluster randomized trial (Moerbeek, 2000)

(if budget = 100.000, cost per cluster = 2000, cost per person = 100)



$$ICC = \frac{variance\ between\ clusters}{total\ variance}$$

Guideline:
Take cluster size for *ICC* halfway realistic range, e.g. 0.05 if range 0 - 0.10
This gives 20 persons per cluster if the cluster/person cost ratio is 20);
The number of clusters then follows from the budget.
If the power is too low then, add budget and clusters, not persons per cluster.

Maastricht University Leading in Learning!

Optimal nr of repeated measures (Winkens, 2005)

	If group difference is linear function of time	If group difference is quadratic function of time
If all correlations equal (compound symmetry)	5 or more, depends on cost ratio (patient : measure)	5 or more, depends on cost ratio (patient : measure)
If correlation decreases as time interval increases (autoregressive)	2 (pre and post only)	3 (pre, halfway, post)

Guideline:
3 measures (pre, halfway, post) is often a good choice,
(but take 5 if much dropout or CS correlations are expected)

Maastricht University Leading in Learning!

Ik heb gezegd